Blucher Mechanical Engineering Proceedings
May 2014, vol. 1 , num. 1
www.proceedings.blucher.com.br/evento/10wccm

10th World Congress on Computational Mechanics
8-13 July 2012 • São Paulo • Brazil

IDENTIFICATION OF DISEASE PATTERNS IN DATA MINING TECHNIQUES

T. M. R. Dias[1], P. M. Dias[1], G. F. Moita[2], P. B. Carolina[3]

[1]Instituto de Ensino Superior e Pesquisa – INESP, Fundação Educacional de Divinópolis – FUNEDI/UEMG

[2]Centro federal de Educação Tecnológica de Minas Gerais – CEFET-MG

[3]Instituto Federal Minas Gerais - IFMG

Abstract. *In the last decades has increased the need for an automated process for discovering interesting information involving large volumes of data. This work is evidence apply data mining techniques to find out the databases of health, patients with Chagas disease, in order to classify the principal characteristics of these individuals. To achieve this goal, will use the free software Weka (Waikato Environment for Knowledge Analysis) to facilitate the process of data mining, along with the Apriori algorithm. To obtain the expected result, during the work will be used a methodology of process of KDD (Knowledge Discovery in Databases), and all phases will be held until the end with a satisfactory result. Having an open source tool, which is responsible for automating the entire process.*

Keywords: *Knowledge Discovery in Databases, Data Mining, Identification of Standards*

Resumo: *Nas últimas décadas, tem aumentado à necessidade de um processo automatizado para a descoberta de informações interessantes envolvendo grandes volumes de dados. Este trabalho tem como evidência aplicar as técnicas de mineração de dados para descobrir em bases de dados da área da saúde pacientes com Doença de Chagas, com o intuito de classificar as principais características destes indivíduos. Para atingir tal objetivo, será utilizado o software livre WEKA (Waikato Environment for Knowledge Analysis) visando facilitar o processo de mineração de dados, juntamente com o algoritmo Apriori. Para se obter o resultado esperado, no decorrer do trabalho, será utilizada uma metodologia de processo de KDD (Knowledge Discovery in Databases), assim serão realizadas todas as fases até chegar ao final com um resultado satisfatório. Tendo uma ferramenta de código aberto, a qual tem a responsabilidade de automatizar todo o processo.*

Palavras Chave: *Descoberta de Conhecimento em Banco de dados, Mineração de dados, Identificação de Padrões.*

1. INTRODUÇÃO

As áreas governamentais, corporativas e científicas têm promovido um crescimento explosivo em seus bancos de dados, superando em muito a usual capacidade de interpretar e examinar estes dados. Esses fatores geram a necessidade de novas ferramentas e técnicas para análise automática e inteligente de bancos de dados [5].

Atualmente, existe grande interesse em desenvolver ferramentas que auxiliem a recuperação de informações eficientes. Diversos esforços de pesquisas têm sido feitos para remediar esse problema.

Segundo Han & Kamber [7], a ampla disponibilidade de imensas bases de dados, aliada à eminente necessidade de transformar tais dados em informação e conhecimento úteis para o suporte à decisão, têm demandado investimentos consideráveis da comunidade científica e da indústria de software. A informação e o conhecimento obtidos podem ser utilizados para diversas aplicações, que vão do gerenciamento de negócios, controle de produção e análise de mercado ao projeto de engenharia e exploração científica.

Pesquisadores motivados pelo desafio de transformar informação em conhecimento, logo se deparam com o Knowledge Discovery in Databases (KDD), dando ênfase na aplicação de mineração de dados.

A mineração de dados ou data mining surgiu em 1989 e consiste de técnicas e algoritmos baseados na análise de padrões para a extração de informações em banco de dados. Sua implementação é dividida em uma sequência de fases, dentre as quais tem-se: a seleção, o pré-processamento, a transformação, a interpretação, a avaliação e a mineração propriamente dita [5] [11].

Várias áreas podem ser beneficiadas com a aplicação da técnica de mineração de dados. Como por exemplo, a descoberta de fraudes por cartões de créditos, a identificação de consumidores nos processos de vendas, as ferramentas de busca na Internet, o apoio nas tomadas de decisões gerenciais, o auxílio no diagnóstico médico em diferentes doenças, entre outros.

Uma vez que há a necessidade da análise de dados referentes à medicina, a proposta deste trabalho baseia-se na utilização de técnicas de mineração de dados para descobrir em bases de dados da área da saúde pacientes que apresentam a Doença de Chagas, com o objetivo de classificar as principais características destes indivíduos.

Dentre as análises realizadas são extraídas informações importantes que possibilitam a tomada de decisão por diversos interessados. Espera-se que os padrões de doenças e áreas de risco sejam automatizados, aumentando, assim, a eficiência de auxílio e suporte a todos os envolvidos.

2. REFERENCIAL TEÓRICO

Nos últimos anos, a quantidade de informação armazenada em base de dados vem crescendo consideravelmente. Com os avanços tecnológicos, os computadores estão cada vez mais rápidos e com maior capacidade de processamento. Além desses fatores, o baixo custo e a automação dos processos de coleta de dados também contribuíram para o aumento de in-

formações armazenadas em base de dados.

Dentro dessa grande quantidade de dados, existem informações preciosas implícitas a nível gerencial e estratégico que não podem ser descobertas pelos sistemas de gerenciamento de base de dados tradicionais.

Uma técnica já bem conhecida e aplicada na análise de grandes bases de dados é a mineração de dados. Fayyad, em 1996, define o conceito de mineração de dados como: “o processo não trivial de identificar em dados alguns padrões válidos, novos, potencialmente úteis e compreensíveis”. Ele também destaca que a aplicação da mineração de dados pode ser inserida em diferentes segmentos e setores da atualidade [5].

O termo mineração de dados é muitas vezes usado como sinônimo de KDD. A mineração de dados é considerada uma etapa essencial do processo de KDD [5][7].

KDD refere-se ao processo global de descoberta de conhecimento a partir de dados, enquanto a mineração de dados é uma fase desse processo.

O KDD baseia-se fortemente em técnicas conhecidas de aprendizado de máquina, de reconhecimento de padrões e de estatísticas para encontrar os padrões nos dados. A estatística oferece, também, métodos de quantificação da incerteza inerente quando se procura inferir padrões gerais a partir de amostras de uma população. As técnicas de visualização de dados estimulam naturalmente a percepção e a inteligência humana, aumentando a capacidade de entendimento e de associação de novos padrões [12].

2.1 Etapas do processo de KDD

O processo de KDD descrito anteriormente contém uma série de fases ou subprocessos definidos, que são a seleção, o pré-processamento e a limpeza, a transformação, a mineração de dados, a interpretação e a avaliação. Essa sequência compreende o ciclo que o dado percorre até se transformar em conhecimento útil, conforme a Figura 1.

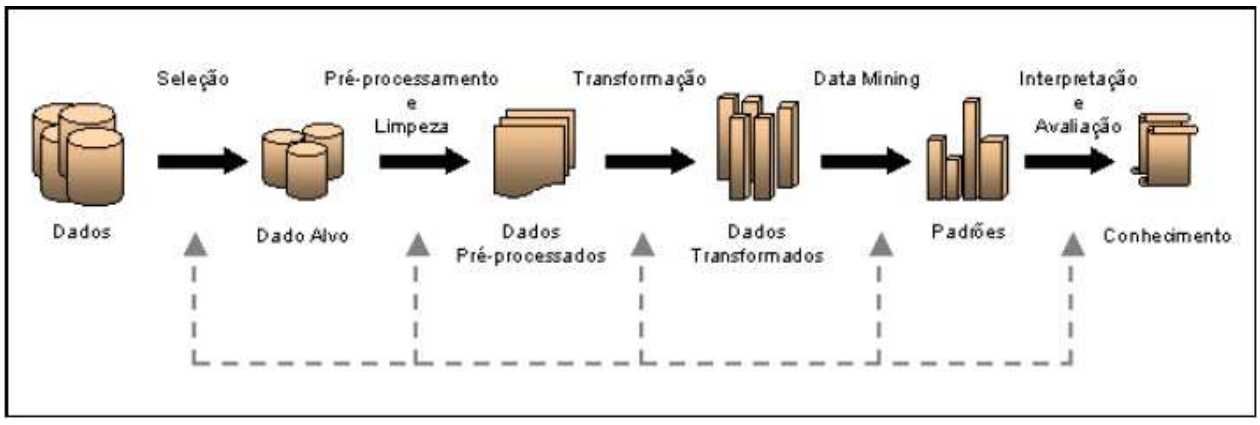

Figura 1. O ciclo do processo de KDD [5].

Inicialmente, Fayyad [5] sugeriu uma proposição para o processo de KDD: “Este processo é composto por uma série de etapas, possuindo uma natureza iterativa e interativa”.

Iterativa, pois, apesar do processo ser composto por etapas sequenciais, pode haver retorno a etapas anteriores; e interativo, pois, neste caso, o usuário poderá optar pela retomada em qualquer uma das etapas deste processo.

A seguir, serão apresentadas cada uma das fases relatadas, para compreensão de suas particularidades.

Seleção:

É realizada uma identificação de quais informações realmente serão trabalhadas na base de dados [6].

Sendo a seleção de um conjunto de dados contendo possíveis variáveis, características e/ou atributos e registros que farão parte da análise.

Segundo Diniz [4], "as variáveis selecionadas para a mineração de dados são denominadas variáveis ativas uma vez que são usadas para distinguir segmentos, fazer predições ou desenvolver outras operações específicas da mineração de dados".

Pré-processamento e limpeza:

É nessa etapa que a qualidade dos dados envolvidos é tratada. Para isso, são realizadas tarefas de limpeza através de um pré-processamento, para extinguir a redundância, a inconsistência e a ausência dos dados, visando adequá-los para a etapa de mineração.

A qualidade é essencial para a obtenção de dados confiáveis. Portanto, dados limpos e compreensíveis são requisitos básicos para o sucesso da mineração [4].

Transformação:

Os dados pré-processados ainda devem passar por uma transformação que os armazenem e os agrupem adequadamente para um formato apropriado para a mineração.

Essa etapa também consiste no pré-processamento final dos dados, removendo os ruídos, tratando os atributos ausentes, padronizando os atributos e removendo os registros repetidos. Após essa etapa, os dados estarão prontos para a realização da mineração. A remoção de ruídos ocorre colocando o valor correto ao dado ou eliminando o mesmo, caso não haja como corrigi-lo. Os ruídos acontecem por diversos fatores no momento da recuperação de diversas fontes, como acidentes (falta de energia elétrica), entre outros. É comum encontrar registros sem valores devido a diversos fatores, como por exemplo: pode acontecer do operador, na entrada dos dados de um cliente, ignorar ou esquecer alguns campos de dados. Dessa forma, é preciso estabelecer critérios para o tratamento dos atributos ausentes.

Mineração:

Todas as etapas do ciclo de processos do KDD devem ocorrer da melhor forma

possível para atingir o resultado desejado. Logo, pode-se dizer que todas as fases do ciclo são importantes para a transformação das informações em conhecimento útil [11].

Entretanto, a fase de mineração de dados pode ser considerada o núcleo do processo de KDD. Ela consiste na busca por padrões de dados. Porém, vale ressaltar que, para que a localização dos padrões ocorra de forma desejada, é necessária a correta realização das fases anteriores.

Para cada problema existe uma técnica ou algoritmo que melhor se encaixa em determinada situação.

Interpretação e avaliação:

Ao final da aplicação da mineração de dados, o padrão de conhecimento obtido é interpretado e avaliado, a fim de verificar novas descobertas e se realmente o seu objetivo principal foi alcançado com sucesso.

A etapa de mineração de dados é muito importante, utiliza técnicas e algoritmos de diferentes áreas do conhecimento. Dessa forma começa com a escolha dos algoritmos a serem aplicados. Essa escolha depende fundamentalmente do objetivo do processo de KDD.

Nota-se que para a realização deste trabalho será utilizado o método de associação, este será detalhado a seguir.

2.2. Regras de associação

A técnica de regras de associação identifica afinidades entre itens de um subconjunto de dados e estas são expressas na forma de regras. Um dos padrões mais comuns que pode ser descoberto a partir do processo de mineração de dados é o conjunto de regras de associação que expressa a probabilidade de um item ocorrer em conjunto. Por exemplo, 80% dos clientes que adquirem o produto “A” também adquirem o produto “B” [13].

Visto que as regras de associação consistem na descoberta de relacionamentos existentes, logo se destaca o Apriori entre vários outros algoritmos que buscam esse objetivo, o qual será descrito abaixo:

Algoritmo Apriori: Miranda et al. [10] afirma que o algoritmo Apriori é considerado um clássico na extração de Regras de Associação. Ele foi proposto pela equipe de pesquisa QUEST, da IBM, que deu origem ao Software Intelligent Miner.

Esse algoritmo procura identificar relações e dependências significativas entre os atributos. Trabalhando com as medidas de confiança e suporte, ele consegue classificar as melhores regras [3].

Esse algoritmo faz recursivas buscas no banco de dados à procura dos conjuntos frequentes (conjuntos que satisfazem um suporte mínimo estabelecido). Possui diversas propriedades que otimizam o seu desempenho, como por exemplo a propriedade de antimonotonia da relação que diz que para um itemset ser frequente todo o seu subconjunto também devem ser, além de utilizar recursos da memória principal e estrutura hash – dados especial, que fazem uma busca rápida e obtém um valor desejado [10].

Segundo Miranda et al [10], a propriedade de Antimonotonia da Relação ou Propriedade Apriori pode ser explicada da seguinte forma:

Se A está contido em B, e A não é frequente, logo B também não é frequente. Isso implica uma diminuição do tempo de execução, pois se A não é frequente, então não será necessário calcular o suporte de B, e o banco de dados não precisará ser varrido.

Logo, quando o assunto é mineração de regras de associação em grandes bancos de dados centralizados, o Apriori é um dos mais conhecidos. Ele pode trabalhar com um número grande de atributos, gerando várias alternativas combinatórias entre eles.

No tópico a seguir uma breve explicação sobre a Doença de Chagas.

2.3. Doença de Chagas

A história de Chagas inicia-se no interior de Minas Gerais. Em abril de 1909, Carlos Chagas (1878-1934), pesquisador do Instituto Oswaldo Cruz (IOC), comunicou ao mundo científico a descoberta de uma nova doença infecciosa humana que acometia operários [8].

Segundo Lannes e Kropf [9], essa doença causada pelo protozoário Tripanossoma cruzi é conhecida como doença de Chagas, em homenagem a quem a descreveu pela primeira vez. O mal de Chagas, como também é chamado, é transmitido, principalmente, por um inseto da subfamília Triatominae, conhecido popularmente como barbeiro. Este animal de hábito noturno alimenta-se, exclusivamente, do sangue de vertebrados endotérmicos. Vivem em frestas de casas de pau-a-pique, camas, colchões, depósitos, ninhos de aves, troncos de árvores, dentre outros locais, sendo que têm preferência por locais próximos à sua fonte de alimento.

3. DESENVOLVIMENTO

Nos últimos anos, com a evolução da tecnologia da informação, a análise dos dados tem sido fator fundamental para várias organizações. Isso ocorre devido ao aumento intenso da quantidade de dados, trazendo a necessidade de extrair informações de maior qualidade e produtividade.

Uma vez que o valor destes dados está ligado à capacidade de gerar informações para tomadas de decisões, setores como o da saúde contam com uma demanda cada vez maior para a análise de dados.

Colazzos [6], descreve que a medicina, valendo-se da evolução da tecnologia da informação, tem gerado uma grande quantidade de dados em uma velocidade cada vez maior. Dessa maneira, surge então a necessidade de ferramentas capazes de analisar e compreender estes dados.

A proposta deste trabalho baseia-se na utilização de técnicas de mineração de dados em uma base de dados referentes a pacientes que apresentam a Doença de Chagas, com o objetivo de classificar as principais características destes indivíduos.

Para se chegar aos objetivos propostos será necessária a aplicação de técnicas de mineração de dados e a utilização de uma metodologia de KDD auxiliada por ferramentas específi-

cas, para assim agregar valor na aplicação da técnica. Sendo assim, o desenvolvimento será por meio da linguagem Java, juntamente com o software Weka, para auxiliar o projeto e dar mais dinamismo e velocidade na aplicação do processo de KDD.

A Figura 2 representa o esquema de desenvolvimento proposto pelo trabalho.

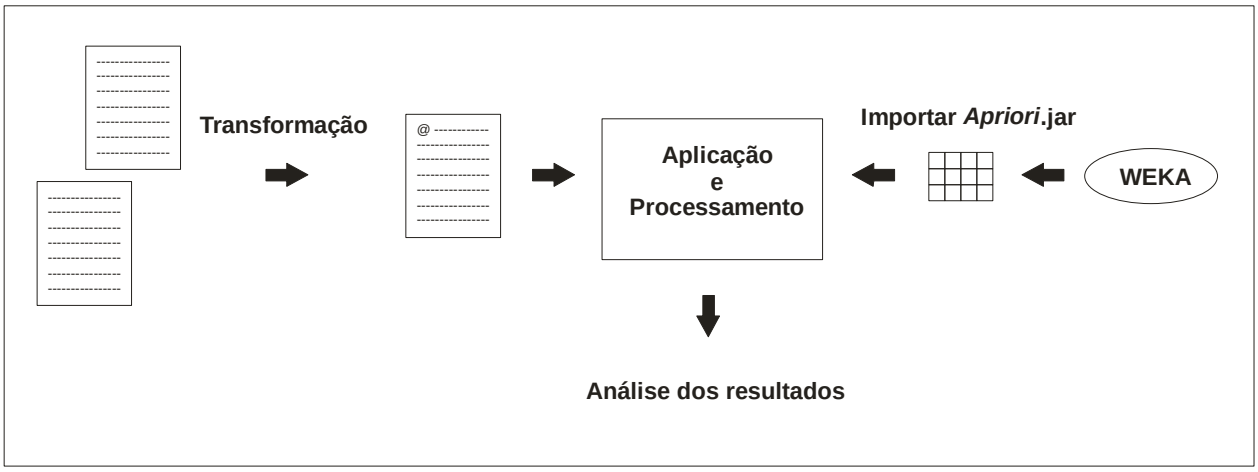

Figura 2. Esquema de desenvolvimento.

Foi desenvolvida uma aplicação através da linguagem Java que fará o registro dos dados coletados, preparados anteriormente, para que estes sofram a mineração e resultem em análise de padrões.

Todos os registros do arquivo serão jogados em um vetor de instância, o qual percorrerá linhas e colunas informando o atributo com seu respectivo valor.

Neste ponto, o arquivo .jar do Weka, especificamente o Apriori.jar, é invocado e o vetor de instâncias juntamente com os parâmetros necessários, sendo informados o número de regras a ser obtido, a confiança e o suporte.

Uma regra é composta de duas partes: uma condição e um resultado. De acordo com Berry [1], o formato geral de associação possui a forma SE “A.”. ENTÃO “B”..., onde “A” é chamado de condição e “B” é chamado de resultado.

Além da condição e do resultado, outros componentes complementam essa regra: suporte e confiança [1]. Berry [1] define o suporte sendo o número de vezes que determinado item ou conjunto de itens aparece em transações distintas relacionadas com o número total de transações operadas, ou seja, suporte A → B: número de registros que contém A e B dividido pelo número total de registros; enquanto que a confiança é o número de transações que suportam o resultado (ENTÃO B...) em relação ao número de transações que suportam a condição (SE A...), ou seja, confiança A → B: número de registros que contém A e B divido por número de registros que contém B.

Com base nisso, o algoritmo Apriori, que faz o cálculo de confiança e suporte, calcula as regras e gera os resultados.

Estes resultados são armazenados e posteriormente retornados em uma string e depois salvo em um arquivo texto.

4. RESULTADOS OBTIDOS

Os resultados alcançados até a conclusão deste trabalho serão apresentados neste tópico sobre a forma de relatórios obtidos pela execução do algoritmo Apriori.

Para se chegar aos objetivos propostos, foi necessária a aplicação de técnicas de mineração de dados utilizando a metodologia de KDD. Sendo assim, o desenvolvimento foi feito por meio da linguagem Java, juntamente com o software Weka, para auxiliar o projeto e dar mais dinamismo e velocidade na aplicação do processo de KDD.

Logo, com a execução do algoritmo Apriori, pôde-se inferir algumas informações a partir das regras que foram obtidas. As possibilidades são muitas, devido à combinação dos atributos que podem ser selecionados e filtrados, como mostra a Tabela 2.

Tabela 1. Resultados da execução do algoritmo Apriori.

Nº Regra	Antecedente	Consequente	Confiança
1	ano=1990 sexo=F 259	cor=Branca 213	0.82
2	ano=1990 492	cor=Branca 400	0.81
3	ano=1990 sexo=M 233	cor=Branca 187	0.8
4	sexo=F	cor=Branca 435	0.8
5	ano=1990 cidade=Bambui 377	cor=Branca 302	0.8
6	sexo=F cidade=Bambui 331	cor=Branca 265	0.8
7	ano=1990 sexo=F cidade=Bambui 198	cor=Branca 158	0.8
8	cidade=Bambui 615	cor=Branca 487	0.79
9	ano=1986 231	cor=Branca 182	0.79
10	sexo=M cidade=Bambui 284	cor=Branca 222	0.78
11	sexo=M 449	cor=Branca 349	0.78
12	ano=1990 sexo=M 233	cidade=Bambui 179	0.77
13	ano=1990 492	cidade=Bambuí 377	0.77
14	ano=1990 sexo=F 259	cidade=Bambui 198	0.76
15	ano=1990 cor=Branca 400	cidade=Bambui 302	0.76

Fazendo a análise dos resultados obtidos, percebemos que a confiança mínima foi decrescendo até chegar a 0.76 (76%), sendo utilizado quatro atributos: sexo, ano, cor, cidade.

Pela regra 1 sabemos que, no ano de 1990, cerca de 259 mulheres de cor branca apresentavam sintomas da doença de chagas, visto que tal regra possui 82% de confiança.

Nesse mesmo ano, foi possível constatar que temos um total de 492 indivíduos, sendo que 400 apresentam cor branca. Isso implica que há 233 homens de cor branca com chagas, de acordo com a regra 2 e tendo cerca de 80% de certeza.

Além disso, percebemos que dos 377 indivíduos que residiam em Bambuí no ano de 1990, e possuíam a doença, 302 tem a cor branca. Percebe-se que o índice de mulheres atingidas é maior que o dos homens - vide na regra 7.

Com um grau de confiança de 78%, no total dos 449 homens, 349 apresentam a cor

branca, sendo que 222 homens de cor branca residiam na cidade de Bambuí, como mostra as regras 10 e 11.

De acordo com a regra 13, em meados do ano de 1990, dos 492 casos de registros de Doença de Chagas foram confirmados 377 em Bambuí, sendo 179 homens e 198 mulheres.

As pessoas brancas são as mais atingidas, visto que, no ano de 1990 há 302 casos, com 76% de confiança, como relata a regra 15.

A compreensão das informações contidas nessas regras é convertida espontaneamente em benefícios significativos para o levantamento de pesquisa com pacientes que apresentam a Doença de Chagas.

5. CONCLUSÃO

Com o crescimento da tecnologia da informação, várias organizações tanto comerciais quanto científicas, mudaram seus métodos de trabalho, devido à facilidade que a computação proporciona no armazenamento de informações.

A informação passou a ser o fator crítico para o sucesso das organizações com o objetivo de conseguir um diferencial e se tornar competitivas. Assim, diversas áreas utilizam técnicas de mineração de dados para alcançar suas metas.

Nota-se que na área da saúde não está sendo diferente. Cada vez mais clínicas, hospitais e órgãos públicos aderem a essa técnica de mineração, buscando encontrar padrões que possam ser utilizados nas tomadas de decisão.

Sendo assim, a proposta deste trabalho baseia-se na utilização de técnicas de mineração de dados em uma base de dados referentes a pacientes que apresentam a Doença de Chagas, com o objetivo de classificar as principais características destes indivíduos.

Por fim, conclui-se que a principal vantagem que esse trabalho trouxe à comunidade acadêmica foi à iniciação no estudo de mineração de dados num ambiente em bases de dados da área da saúde, com pacientes que apresentam a Doença de Chagas, apresentando e exemplificando as principais ferramentas e técnicas. Este, ainda, permite abrir frente para novas e interessantes pesquisas. Uma ferramenta deste tipo contribui significantemente em qualquer que seja o segmento utilizado.

6. REFERÊNCIAS

[1] BERRY, M. J. A.; LINOFF, G. (1997) “Data Mining Techniques: for Marketing, Sales and Customer Support.”, New York: John Wiley & Sons, 454p.

[2] COLLAZOS, K.S.; Barreto, J.M.; Roisenberg, M. (2002) “Dificuldades na aplicação de KDD em medicina.”. Disponível em: <http://200.169.53.89/download/cd%20congressos /2002/2%.

[3] CORRÊA, K. S. (2009) Processo de mineração de dados no estudo de fenômenos solares e geomagnéticos. 28 f. Trabalho Acadêmico (mestre) – INEP, São José dos Campos.

[4] DINIZ, C. A. R.; LOUZADA NETO, F. (2000) “Data mining: uma introdução.”, São Paulo: ABE, 123p.

[5] FAYYAD, U; SHAPIRO, G. P; SMYTH, P; UTHURUSAMY, R. (1996) Advances in Knowledge discovery and data mining.”, MenloPark: Mit Press, 560P.

[6] GOLDSCHMIDT, R.; PASSOS, E. (2005) “Data Mining um guia pratico.”, 1. ed. Rio de Janeiro: Campus.

[7] HAN, J.; KAMBER, M. (2001) “Data Mining: Concepts and Techniques.”, San Francisco: Morgan Kaufmann Publishers.

[8] KROPF, S. P. (2007) “História da doença de Chagas: ciência, saúde e sociedade”, Rio de Janeiro, RJ, Brasil, Disponível em: http://www.fiocruz.br/chagas/cgi/cgilua.exe/sys/start.htm?sid=171. Acesso em: 28 de outubro de 2010, 14:32.

[9] LANNES, J.; KROPF, S. (2007) “A doença”, Rio de Janeiro, RJ, Brasil, Disponível em: http://www.fiocruz.br/chagas/cgi/cgilua.exe/sys/start.htm?sid=91**,** Acesso em 28 outrubro de 2010, 14:37.

[10] MIRANDA, D.; SABORÊDO, A. P., et al. (2003) “Iniciação Científica – Data Mining.” AEDB Associação Educacional Dom Bosco. Resende - Rio de Janeiro.

[11] PRASS, F. S. (2004) “KKD: Processo de descoberta de conhecimento em bancos de dados.”, Grupo de Interesse Em Engenharia de Software, Florianópolis, v. 1, p. 10-14.

[12] REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; PAULA, M. F. (2002) “Sistemas inteligentes: fundamentos e aplicações de Mineração de dados.”, In: REZENDES, S. O. (Org.), Barueri: Editora Manole, p. 307-335.

[13] ZANDER, V. P. S. (2007) “Mineração de Dados – Conceito, Aplicações e Experimentos com Weka.”, Disponível em:<hhtp://www.sbc.org.br/bibliotecadigital/download.php ?paper=35>. Acesso em: 20 agosto 2009, 17:22:10.